A classroom observation tool for scaffolding reading comprehension

Smit, Nienke; van de Grift, Wim; de Bot, Kees; Jansen, Ellen

Published in:
System

DOI:
10.1016/j.system.2016.12.014

Publication date:
2017

System 65 (2017)

A classroom observation tool for scaffolding reading comprehension

Nienke Smit a, *, Wim van de Grift a, Kees de Bot b, Ellen Jansen a
a Department of Teacher Education, University of Groningen, The Netherlands
b Department of Applied Linguistics, University of Groningen, The Netherlands

Keywords: Scaffolding, observation, teacher education, reading, generalizability study, content-based teaching, EFL

Abstract

An important goal of educational research is to find out which teaching practices are effective in promoting students' learning. In order to assess these practices, adequate observation instruments are needed. Existing observation schemes for language teaching are not suitable to gauge which teaching strategies scaffold EFL reading comprehension in particular and language learning in general. Therefore, we developed a new instrument: the English Reading Comprehension Observation Protocol. The focus of the instrument is on the role of the EFL teacher who helps students to move from learning to read to reading to learn in English. We conducted a generalizability study in order to establish the instrument's reliability. Twenty lessons taught by five experienced teachers were recorded and observed by five experienced teacher educators. The results of the generalizability study, in which we disentangled sources of variance, show that a large proportion of the variance can be attributed to differences between the teachers. This shows that the instrument has a high reliability and can help teachers identify their strengths and room for development. The instrument takes the form of a checklist and is easy to use for professional development purposes.

* Corresponding author. Department of Teacher Education, University of Groningen, Grote Kruisstraat 2/1, 9712TS Groningen, The Netherlands. E-mail address: N.Smit@rug.nl (N. Smit).

System 65 (2017)

1. Introduction

In a rapidly changing and globalized world in which large amounts of written information need to be processed and understood, reading to learn in English is a crucial skill when preparing teenagers to work, study and live in diverse contexts (Grabe & Stoller, 2011; OECD, 2016; Walqui & Van Lier, 2010), More than 25 years ago, Snow, Met, and Genesee (1989) proposed a conceptual framework for integrating language and content teaching in second and foreign language classrooms. They made a plea for language teachers to incorporate "meaningful and important content that has evident language related value in the rest of the curriculum" (p. 213). This comprises using English to learn about content outside the domain of language and goes beyond learning to use English in communicative situations (Coyle, Hood, & Marsh, 2010; Richards & Rodgers, 2014; Lightbrown, 2014).

The role of the EFL teacher in a content-based approach or a content and language integrated learning (CLIL) program is to help students improve their English reading comprehension skills, not only to learn the language but more so to help the learner understand and evaluate the meaning of a text (Juan-Garau & Jacob, 2015). According to Coyle et al. (2010) and Gibbons (2002) reading comprehension involves: building on existing knowledge, learning the target language, learning through the target language, learning about the target language, and integrating meaning into new contexts. Many teachers in the Netherlands and elsewhere lack the skills and instruments to assess the individual critical literacy levels of students (OECD, 2016). Instructed language learning can assist foreign language acquisition in general and developing the complex competence of reading comprehension in particular (Grabe, 2009; Long & Doughty, 2009).

In order to better understand how instructed language learning takes shape in authentic classrooms, observational studies are needed. Classroom observations provide rich empirical data about what is going on in classrooms and can give us more information about the role of the EFL teacher in the process of developing skilled and critical second language readers. The English Reading Comprehension Observation (ERCOP) instrument we are presenting in this paper attempts to observe and quantify the support an EFL teacher gives their students.

2. Literature review

2.1. Strategies for scaffolding reading comprehension

In order to be able to support a whole class with calibrated and adaptive support, we contend that teachers need a range of strategies, which they can use consistently and flexibly in their EFL teaching. The concept of scaffolding forms the underlying theoretical basis for the observation instrument. Therefore, we will first define the concept of scaffolding.

Giving learners calibrated and appropriate support is often referred to as scaffolding. Wood, Bruner and Ross (1976) explored the concept in tutorial one-to-one interactions and introduced the term scaffolding. In their analysis of the tutoring process they speak about six "scaffolding functions": recruitment of interest, reduction in degrees of freedom, direction maintenance, marking critical features, frustration control, and demonstration. Nuttall (2005) indicates that scaffolding is something some teachers do intuitively, but that "other teachers have to learn how to do it" (p.36).

In this paper, scaffolding is defined as the temporary and contingent teacher support that helps learners to comprehend a text, to carry out a reading comprehension task and to produce meaningful output in a second or foreign language. (Gibbons, 2002; Walqui & Van Lier, 2010; Wood et al., 1976). Reasons for using scaffolding strategies in EFL reading lessons include supporting metacognitive activities, cognitive activities, student affect, and fostering engagement. Although scaffolding is interactive in nature and dependent on the teaching context, common elements in existing research can be pinpointed (Van Geert & Steenbeek, 2005; Van de Pol, Volman, & Beishuizen, 2010).

The first element is the use of diagnostic strategies and contingent support (i.e. support that is adjusted to the current and prospective level of a student's learning).Walqui and Van Lier (2010) call this kind of foreshadowing “proleptic interactions” (p.24).

The second element of scaffolding is fading: withdrawing support when appropriate. According to Nuttall (2005), this entails “never doing anything [for your learners that] they are capable of doing for themselves with a little support” (p.36). She also proposes that teachers, who often have limited time available for individual students, use the steps individual students take as opportunities for everyone to learn. This can be done in oral interaction through dialogic teaching, but not in a classroom dominated by a transmission-style of teaching.

The third common element in scaffolding research is transfer of responsibility from teacher to learner in the learning process (Foley, 1994; Van de Pol et al., 2010). This should be understood in terms of ensuring that students understand their own learning and enabling students to take charge of this process in order to become independent and successful readers (Nuttall, 2005).

Tharp and Gallimore (1988) identified the following “means of assisting performance” strategies (scaffolding means): modeling, contingency management, feeding back, instructing, questioning, and cognitive structuring. Based on Tharp and Gallimore and Wood et al. (1976), Van de Pol et al. (2010) created a framework for analysis of scaffolding strategies with five scaffolding intentions and six scaffolding means. Their framework distinguishes scaffolding goals or intentions and scaffolding means. Intentions for scaffolding indicate underlying reasons for scaffolding (i.e. why) and focus on what aspect of learning is scaffolded (metacognitive, cognitive or affective processes), scaffolding means, on the other hand, focus on how learning is being scaffolded.

Reasons for using scaffolding strategies (scaffolding intentions) cannot be inferred from observational data, but teaching strategies (scaffolding means) can be operationalized in an observation tool for reading comprehension.

2.2. Strategies for scaffolding reading comprehension

Even though research into second language acquisition asserts that exposure to meaningful input is essential for language learning, comprehensible input alone is not enough. Input should not be simplified and impoverished, but elaborate and rich (Long & Doughty, 2009). We will first focus on the role of the teacher in fostering reading comprehension development.

There is ample evidence that instruction is most effective if it combines attention to form and meaning (rather than a focus on one of them) (Norris & Ortega, 2000, 2006; Spada & Tomita, 2010). The teacher can facilitate language development, but only if the learner is willing to make a cognitive contribution to

the acquisition process (Long & Doughty, 2009; Lyster, 2007). Engaged and active students are more likely to make a cognitive contribution. Many studies of teaching reading comprehension stress the importance of activating students' background knowledge in interaction with teacher and text by means of hands-on tasks (Grabe, 2009; Grabe & Stoller, 2011; Guthrie, Wigfield, & Perencevich, 2004; Nuttall, 2005).

Activating a classroom full of teenage students is complex (Van de Grift, Helms-Lorenz, & Maulana, 2014). However, meta-analyses by Hattie (2009) revealed that effective teachers know how to integrate "knowledge, empathy and verbal ability" (p.115) and combine, change and add to the lessons according to their students' needs and their teaching goals. Hattie also states that exposure to information alone is not enough to foster learning. Effective teachers do not facilitate learning, they activate learning. Examples of teachers as activators are found in studies into reciprocal teaching, feedback, teaching students self-verbalization, direct instruction, meta-cognitive instruction, providing worked examples (modeling a task in order to reduce learners' cognitive load), providing goals and behavioral organizers (Hattie, 2012). Expert teachers act contingently, and contingent teaching is one of the key elements of scaffolding (Van de Pol et al., 2010). Hattie's conclusions resonate with what Grabe (2009) notes about combining content and reading instruction. Content-based reading instruction offers "many natural opportunities for (a) extended reading; (b) motivational learning experiences; (c) strategic responses to complex tasks; (d) greater choices in reading materials; and (e) matching growing reading skills with more challenging tasks" (Grabe, 2009). Teacher training and teacher development are needed when schools want to implement content-based reading.

Content-based reading encourages a formative approach to learning, also known as assessment for learning (Assessment Reform Group, 2002). This approach focuses on development. Assessment for learning helps learners develop effective learning strategies and motivates learners to improve (Grabe, 2009). In addition, assessment for learning gives both teacher and students information about student performance and progress in the reading classroom (Van Gorp & Van den Branden, 2015). Strategies that foster this kind of diagnostic information include student self-evaluation, asking students about the reading process and progress, and asking students to generate questions about the text (Coyle et al., 2010; Gibbons, 2002; Grabe, 2009; Walqui, 2006). These strategies can be subsumed under the notion of scaffolding.

2.3. Existing observation tools

Observation tools can be helpful in evaluating the effectiveness of professional development programs and in developing teacher education programs (Desimone, 2009; Hill, Charalambous, & Kraft, 2012). For our study, we need an observation instrument that captures a range of teaching strategies to scaffold reading to learn. Developing such an instrument is an intricate process: the categories of observation need to have an empirical and theoretical foundation and the observations need to lead to comparable results from different observers, Additionally, the instrument needs to be easy to use for teachers and teacher educators.

There are several established observation instruments for the communicative orientation of L2 classrooms – COLT (Fr€ohlich, Spada, & Allen, 1985), motivation in the language classroom e MOLT (Guilloteaux & Dörnyei, 2008), sheltered instruction in immersion education e SIOP (Echevarria, Vogt, & Short, 2010) and for subject teachers in a CLIL program (De Graaff, Koopman, Anikina, &Westhoff, 2007). However,

these instruments are not specific for teaching scaffolding strategies in the EFL reading comprehension classroom: they were designed for a different purpose and either do not mention the term scaffolding (Fröhlich et al., 1985) or mention scaffolding as a general and commonly understood strategy a teacher could and should use (De Graaff et al., 2007; Echevarria et al., 2010; Guilloteaux & Dörnyei, 2008).

We will briefly discuss these four existing observation instruments here, thereby focusing on scaffolding. The Communicative Orientation of Language Teaching scheme (COLT) focuses on capturing differences in communicative orientation of L2 classrooms (Fröhlich et al., 1985). Scaffolding and reading comprehension are not explicitly part of this scheme.

The Motivation Orientation of Language Teaching scheme (MOLT) focuses on motivational strategies for language teaching (Guilloteaux & Dörnyei, 2008). This scheme describes scaffolding as "providing appropriate strategies and/or models to help students complete an activity successfully (e.g. the teacher thinks aloud while demonstrating, reminds students of previously learned knowledge or skills that will help them complete the activity, or has the class brainstorm a list of strategies to carry out the activity)" (Guilloteaux & Dörnyei, 2008). It is unclear when a strategy is deemed appropriate, and what kind of modeling or thinking aloud helps students to complete an activity successfully.

De Graaff et al. (2007) developed an instrument for subject teachers in Content and Language Integrated Learning (CLIL) that focuses on language teaching outside language lessons. Scaffolding is mentioned explicitly in this instrument as an indicator in the scale "teacher facilitates the use of strategies" (p.609e610). They consider scaffolding to be very important, however, this indicator requires high-inference from the observers. It is unclear how scaffolding can be identified reliably.

The Sheltered Instruction Observation Protocol was designed for observations of sheltered instruction. Sheltered instructed caters for the needs of English language learners in an immersion context (Echevarria et al., 2010). Out of 30 observation points, only item 14 mentions scaffolding as techniques that should be "consistently used assisting and supporting student understanding (e.g. think-alouds)" (p.229). In order to operationalize scaffolding for classroom observation, clarity about which teaching behaviors are qualified as assistance and support is needed.

SIOP, MOLT and the observation protocol for CLIL teachers have included scaffolding as one discrete observation point. As we have seen in the literature review (section 2.1.), scaffolding is too complex to capture in one concrete observation point. In order to identify and observe teaching strategies for scaffolding reading comprehension, we will operationalize the six different scaffolding means.

3. The study

3.1. Research question

In the present study we view scaffolding as an advanced teaching skill, which consists of six different aspects (giving instructions, explanations, hints and feedback, modeling and asking questions). Because this teaching skill cannot be observed in one concrete observation point, the following two questions guide the present study:

1. How can we develop a reliable and practical observation protocol for scaffolding EFL reading comprehension?

2. How acceptable is the protocol's reliability when it is used by a small number of observers?

3.2. Context: EFL in Dutch secondary education

We conducted this project in both regular schools and schools with bilingual streams. In all Dutch secondary schools (bilingual and mainstream) English is a core curriculum subject. Communicative language teaching principles dominate secondary classrooms in the Netherlands (Stichting Leerplan Ontwikkeling, 2013). The curriculum for modern foreign languages is organized around the Common European Framework of Reference domains listening, reading, speaking (production and interaction) and writing, supplemented with literature (Meijer & Fasoglio, 2007).

Fifty percent of the grade for English is determined by a national reading comprehension exam at B2/C1. This means that the average student should be able to read complex texts about any topic with low frequency words, a wide range of vocabulary and consisting of long and complex sentences when s/he enters university (Council of Europe, 2001). It is one of the tasks of English teachers to help learners to read English texts efficiently whilst paying attention to the content of the text (Meijer & Fasoglio, 2007; Stichting Leerplan Ontwikkeling, 2013).

Most teaching materials focus primarily on training reading skills and ensuring that students are capable of finding the correct answers to pre-specified (mostly multiple-choice) text comprehension questions. A salient type of reading activity in the Dutch EFL classroom is the formal or informal assessment of reading comprehension through post-reading questions.

However, it is doubtful whether or not this teaches the students to derive meaning from a text, or provides sufficient evidence of the students' understanding of a text (Nuttall, 2005).

Except for bilingual schools in the Netherlands, where CLIL and cross-curricular collaboration is required, an approach that integrates a focus on content and language in a language classroom is not common practice in mainstream Dutch schools.

3.3. The instrument: creating the observation tool

We were specifically interested in designing a practical tool in which we took the teacher in the role as activator of cognitively challenging interaction as a core criterion. We wanted to develop a reliable observation instrument that gives as much information as possible about the teachers (object of measurement in our study). The tool should be able to give the observed teacher concrete feedback about what they are already doing in terms of scaffolding strategies and to establish room for expansion of their flexible teaching toolkit. Two of the most frequently used sampling methods in observational research are time sampling and event sampling. Time sampling was considered to be less useful as our focus is on occurrences of complex teaching strategies. We are interested in learning whether teachers show these strategies or not. Structured observation in event sampling enables us to make comparisons between lessons (Cohen, Manion, Morrison, & Bell, 2011; Shavelson & Webb, 1991; Spada & Lyster, 1997). Inorder to be able to compare across different lessons and teachers, event sampling was chosen (Mackey & Gass, 2012).

As a starting point for our list of events, we used the six scaffolding means as identified in the framework for analysis of scaffolding strategies by Van de Pol et al. (2010): feedback, hints, instructing, explaining,

modeling and questioning. Drawing on the L2 teaching literature, we compiled a list of 71 items consisting of possible teaching strategies (Dalton-Puffer, 2007; Echevarria et al., 2010; Gibbons, 2002; Grabe, 2009; Grabe & Stoller, 2011; Guthrie et al., 2004; Lyster & Ranta, 1997; Lyster, 2007; Walqui & Van Lier, 2010).

The English Reading Comprehension Observation Protocol (ERCOP) is a list of possible strategies and was designed as a checklist in which each item was marked as observed or not. The observation instrument for general teaching strategies, the International Comparative Analysis of Learning and Teaching (ICALT) observation instrument (Van de Grift, 2007, 2009) was used as a model. To avoid the central tendency, we used a four-point Likert scale. Score 1 carries the label "weak" and implies that the teacher does not use the strategy or could have used the strategy but fails to do so (a missed opportunity). Score 2 means "weakness rather than strength" indicating that the teacher only uses the strategy occasionally. Scores 1 and 2 express a negative verdict. Score 3, "strength rather than weakness", indicates that the teacher sometimes uses the strategy, and score 4 "strength" that the teacher consistently uses the strategy (often/always). Scores 3 and 4 express a positive verdict. The observers also had the option to indicate on the observation form that the behavior was not applicable or not observed, but they were strongly discouraged to use this option. The option was only applicable if they felt that none of the other scores applied, for example, if a teacher could not have used it because the strategy is irrelevant for the lesson In order to receive feedback on the clarity and comprehensibility of the 71 items, we tried out this observation list with 99 observers (MA TESOL students (n ¼ 93) and teacher educators (n ¼ 6)) on one lesson. A video of an EFL lesson to a group of 15- year-old students in a mainstream Dutch secondary school was used for the trial version of the instrument.

The 99 observers reported that scoring 71 items while observing a 30-min lessonwas too demanding in terms of cognitive load. Based on their responses, the list was reduced to 45 observation points, which were deemed achievable in terms of cognitive load. Items were deleted or merged for the following four reasons. First of all, observers had to make high inferences about what was intended, which was the case for the item "helps the student to bridge the gap between everyday language and academic language". Secondly, some items could not be observed, for instance "modifies/replaces the textbook texts".Thirdly, items, which capture teacher-student interactions that take place over a longer period of time, cannot be scored on a rating scale. For example "helps students to self-verbalize the steps to complete the reading task". Lastly, some of the observation points were modified in order to make them more transparent. The observations points "helps students to give oral output during group work; helps students to give oral output during individual work; helps students to give written output during group work; helps students to give written output during individual work" were merged into "helps students to give L2 output".

We used this new version of the ERCOP instrument, containing 45 items, to train five observers. In the next section we will outline the observer training.

3.4. Training the observers

The observers received a three-hour training by the first author of this article to make sure that they were all looking through the lens of the structured observation protocol. During the training they were informed about the background of the instrument. The aims of the training were (1) to ensure a shared understanding of the observation points and (2) to enhance reliability. It was pointed out to the observers

that the goal was not to reach consensus regarding the observers' beliefs of what constitutes good EFL teaching, but rather to reach consensus in their observations with this protocol.

We set out to achieve these training aims by studying the items together, focusing on the meaning of the items and instilling the meaning of the items through discussion. The discussions were important to raise the observers' awareness of their personal preferences. The observers were trained not to make judgments based on their personal preferences and to repress their intuitive verdicts. They learnt how to focus on the observation points and the meaning of the scores. During the training the observers watched and rated two 10-min videos that were not part of the data set used for the generalizability study (see sections 3.7 and 3.8). As soon as the observers had filled out the list, the trainer and a research assistant collected the scores and entered them into a data file in SPSS. The level of consensus was calculated by transposing the scores of a four-point scale into dichotomous scores. We used the dichotomous scores to calculate the percentage of agreement between the observers for every item (70% was required). The trainer discussed the items causing debate (<70% agreement) with the observers. The observers were asked to explain how they established their score. After the discussion the observers were given the opportunity to alter their initial response. This exercise was done twice during the training session.

At the end of the training session there were four items (out of n ¼ 45) on which the observers still disagreed. We calculated the intra class coefficient to assess absolute agreement between raters. For every individual rater, we found a coefficient of 0.61 for the first video and a coefficient of 0.65 the second video. This was considered sufficient agreement to start the data collection.

3.5. Purpose of the generalizability study

For the development of observation protocols, Hill et al. (2012) make a plea for using a generalization study (G-study), which is a relatively unknown method in L2 teaching research (Derrick, 2015). Therefore, it will be elaborated upon here. Factors that impact observational research include "when, where and for how long we look, how many observers there are, and how we look." (Cohen et al., 2011). The reliability of observational data can be compromised by the moment of observation, the group of students that is being observed and the influence of the observer. Generalizability theory's (G-theory) premise is that an observation or measurement of a person is no more and no less than one random sample of that person's behavior. G-theory enables us to isolate sources of variation (facets in G-theory) in measurement such as observer, group, moment, teacher and item. An important advantage of using generalizability theory is that it can disentangle multiple sources of measurement variance in one analysis (Shavelson & Webb, 1991). In G-theory it is assumed that a person's measured attributes are in a steady state. G-theory focuses on the magnitude of sources of variance that are affecting the measurement and it allows us to estimate the magnitude of the sources of variation using analysis of variance (ANOVA).

In the current G-study, we assume that any differences in teachers' scores are not the result of systematic changes in the teacher, but for instance the result of potential inconsistencies in observers' observations. These sources of variance might flaw the reliability of the results. Some observers might be more lenient or stringent than others. Similarly, the group atmosphere or a particular time of day might affect the flow or teaching performance of the teacher.

3.6. Participants

The team of teachers consisted of five secondary EFL teachers recruited from the network of the first author who voluntarily agreed to participate in the study. The sample was a nonprobability sample, because a prerequisite for inclusion was that teachers were skillful classroom managers. The teachers (two males, three females with ages ranging from the mid-thirties to early sixties) were all senior teachers with more than 10 years of teaching experience and worked at four different schools. All teachers hold a Masters' degree in English (see Table 1). The teachers were asked to teach business-as-usual lessons. The only inclusion criterion for the lessons was that reading comprehension was part of the lesson. Teaching materials were not prescribed. CLIL teachers taught their lessons entirely in English. In mainstream lessons, students and teachers occasionally used the L1 (Dutch).

Table 1: Participants' Background Information

	Gender	Age	L1	Workplace
Anna (T1)	Female	Late 30s	Dutch	CLIL school
Bob (T2)	Male	Mid 30s	Dutch	CLIL school
Charlotte (T3)	Female	Early 40s	Dutch	Mainstream education
David (T4)	Male	Mid 50s	Dutch	Mainstream education
Emma (T5)	Female	Early 60s	English	CLIL school

The five secondary EFL teachers were video recorded four times each: in two different groups of students at two different moments in time. The class size for all ten classes was typical for standard secondary school groups (between 23 and 31 students per group) in the Netherlands. This resulted in a dataset of 20 lessons. Eighteen lessons were filmed in upper school groups (students aged 15e18), two lessons were filmed in lower school groups.

The original recordings of the 20 lessons lasted between 43 and 60 min per lesson. In order to make the observations manageable in terms of concentration and cognitive load, the lessons were edited to fragments lasting between 13 and 20 min. All instances in which EFL teaching strategies and teacher-student interaction were visible were included in the fragments, but seat work, quiet reading time and procedural activities were shown only for a few seconds to give observers an impression of the flow of the lesson. These phases were not coded in the observation protocol.

The team of trained observers was made up of five modern foreign language teacher educators (four English, one German[1]). The observers were selected on their professional experience in making informed judgments about teachers.1 Background information about the observers can be found in Table 2 below.

[1] The Dutch curriculum for German and English is based on the same attainment targets (Meijer & Fasoglio, 2007). Although she originally has a teaching background in a different foreign language (German), this teacher educator is also experienced in educating and observing EFL (trainee) teachers.

Table 2: Observers' Background Information

	Gender	Age	L1	Job title
Fiona (O1)	Female	Early 30s	Dutch	EFL teacher educator
Gijs (O2)	Male	Early 60s	Dutch	EFL teacher educator
Harold (O3)	Male	Early 30s	Dutch	EFL teacher educator
Iris (O4)	Female	Early 50s	Dutch	EFL teacher educator
Jenny (O5)	Female	Early 40s	Dutch	German FL teacher educator

3.7. Method: Generalizability theory

The goal of our observations was to capture the range of strategies the EFL teacher uses during the lesson. Observations in this study took the form of non-participant, pre-specified and structured observation in a natural setting. The design of this G-study (see Fig. 1) is a two-facet partially mixed design, in which the lessons (l) are nested within teachers (t) and crossed with observers (o). This is abbreviated as (l:t) o. All raters coded all teachers on all occasions (lessons in this study) (Shavelson & Webb, 1991). Because some facets are confounded in this design, not all sources of variability can be estimated separately. The lesson effect is confounded with the teacher-by-lesson interaction. The observer-by-lesson interaction (ol) is confounded with the three-way interaction and remaining unmeasured sources of variance (tol,e).

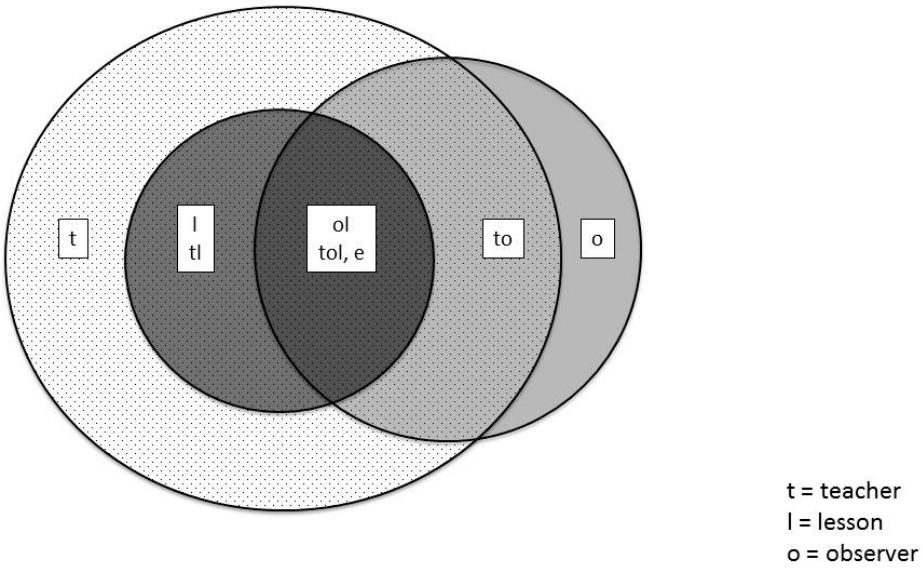


Figure 1 Generalizability study Venn diagram: a two facet-partially mixed design (l:t) x o

The facets (sources of variance) that can be distinguished in our study are teacher (t), lesson (l:t), observer (o) and the interaction between teacher and observer (t x o). The area of the Venn diagram (Fig. 1), where three circles intersect, comprises the interaction between observer and lessons (o x l), the interaction between teacher, observer, and lessons (t x o x l), and a residual (e, for unsystematic error).

3.8. Scoring the data set

All observers received the exact same dataset with 20 video-recorded lesson fragments. They received the videos in a different and random order to level out boredom, fatigue and recency effects. To avoid biased memories, the observers were asked to fill in the scores on the 45-item instrument while watching the videos. The observers had 4 weeks to submit the completed observation lists.

Written instructions were given informing observers of the aim of the study: to find out if the strategies listed on the protocol could be observed reliably. It was pointed out that the observation points were not listed in chronological order and that all observation points were categorized in terms of the following scaffolding means: instruction, explanation, feedback, hints, modeling, questioning. The instructions also alerted observers to the fact that they might observe behaviors that were not listed on the protocol and they were asked to ignore them. Observers were asked to stop scoring fragments when starting to feel tired or unfocused. Finally, observers were reminded that they were trained to look through the lens of the protocol and should ignore personal opinions or preferences as much as possible.

The observers scored the observation points on the 4-point Likert scale (same scale as the pilot phase). The observers had the option to indicate on the observation form that the behavior was not applicable or not observed, but they were strongly discouraged to use this option. The option was only applicable if they felt that none of the other scores applied, for example, if a teacher could not have used it because the strategy is irrelevant for the lesson.

4. Results and discussion

4.1. Reliability of the instrument

After the five observers had all scored their set of 20 lessons, the G-study analysis was performed. We used the MIXED maximum likelihood model procedure with fixed facets in SPSS to partition the sources of variance in order to establish the magnitude of the following sources of variance: teachers, observers, lessons, teachers crossed with observers and residual. In the first data analysis we found three items that caused an error (Hessian matrix was not positive) in SPSS. The following three items were reconsidered and dropped:

- amplifies but does NOT simplify language (hints

- arouses curiosity in the text (modeling)

- creates vivid mental images that help students to better understand the text (modeling)

The fourth item that was deleted is “has clear reasons and goals for an activity”. All four items were too difficult to observe. We considered words and phrases such as “amplify”, “simplify”, “arousing curiosity” and “creates vivid mental images” to be too subjective.

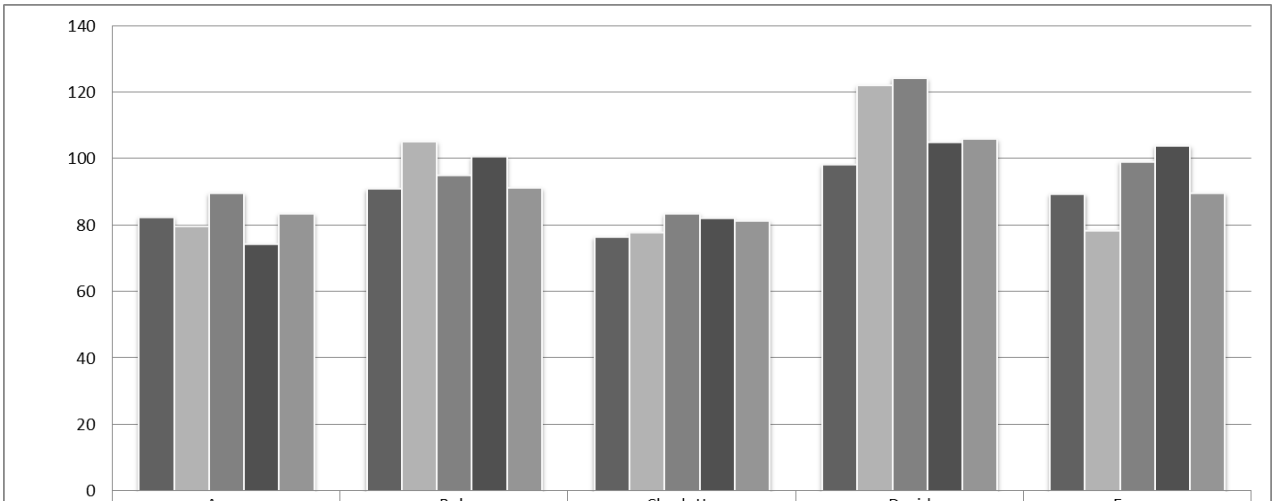
A new G-study analysis with a list of 41 observation points was conducted. Table 3 shows the results of the G-study analysis. The variance components (main effects) are presented in Table 3 below.

Table 3 Variance decomposition for ERCOP scale

	% Variance ERCOP scale
Teacher (t)	51.67
Observer (o)	1.99
Lesson (l:t)	1.06
Teacher x observer (to)	18.14
Residual	27.13
Total	99.99

Note: cells represent the percentage of variance explained by different facets in a generalizability study. The total is 99.99 due to rounding.

Because the aim is to have a reliable instrument that is sensitive to differences between the observed teachers, and insensitive to influences caused by observers, an unruly class or a Friday afternoon lesson, the protocol should find a low proportion of variance caused by lessons and observers and a high proportion of variance caused by teachers, Table 3 shows that 51.67% of the total variance can be attributed to observed differences in teachers. The remaining 48.33% can also be accounted for. The influence of the observers was very low (1.99%), suggesting that the observers are interchangeable. The percentage of variance for lessons was also very low (1.06%). There was a higher, but still relatively low, percentage of the variance in the observations (18.14%) caused by interaction effects between teachers and observers. This means that some observers had a personal preference for a particular teacher. The differences between the observers are illustrated in Fig. 2 below.



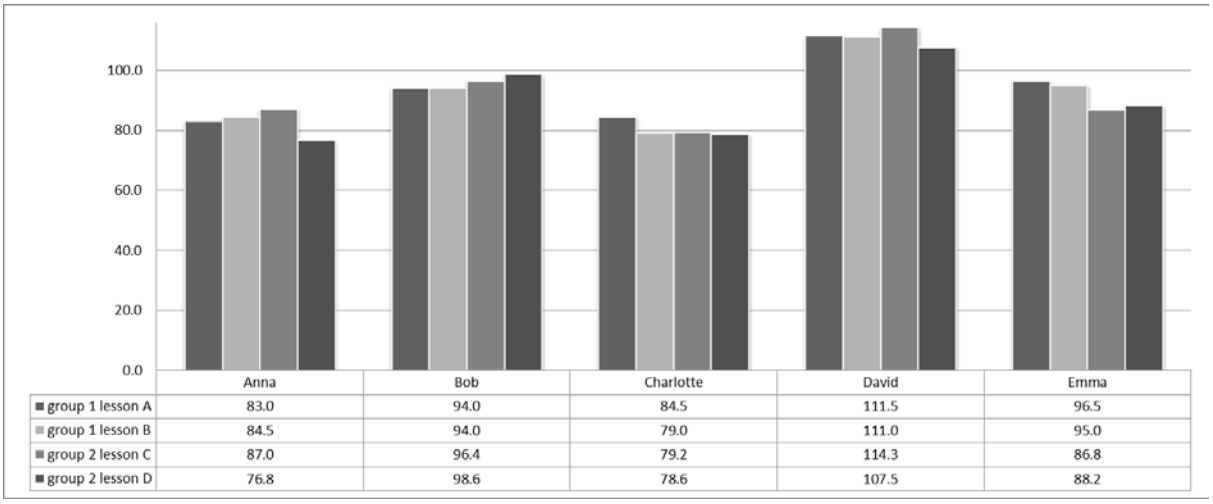
	Anna	Bob	Charlotte	David	Emma
Fiona	82.5	91	76.5	98.25	89.5
Gijs	79.75	105.33	77.75	122.25	78.5
Harold	89.75	95	83.5	124.33	99.25
Iris	74.33	100.67	82.25	105	104
Jenny	83.5	91.25	81.5	106	89.67

Figure 2 Mean observer scores on the ERCOP scale per teacher

From Fig. 2 we can see that one observer, the senior male teacher educator, was slightly more positive about the two male teachers and relatively less positive about the female teachers.

Some differences between the lessons are visible in the scores. Teachers Anna and Charlotte used reading comprehension exams (past papers) as teaching materials in their lessons, whereas the other three teachers used their own materials. Anna and Charlotte were using assessment of learning materials for assessment for learning purposes.

In her second lesson, one teacher (Charlotte) became very angry with a group of 17-year-old students who had failed to do their homework. As a consequence, a substantial proportion of lesson timewas spent on silent reading in order to catch up for the missed homework. This obviously resulted in a lower score on the ERCOP scale, because fewer teaching strategies could be used. Fig. 3 below, illustrates these results.



	Anna	Bob	Charlotte	David	Emma
group 1 lesson A	83.0	94.0	84.5	111.5	96.5
group 1 lesson B	84.5	94.0	79.0	111.0	95.0
group 2 lesson C	87.0	96.4	79.2	114.3	86.8
group 2 lesson D	76.8	98.6	78.6	107.5	88.2

Figure 3 Mean ERCOP score per lesson

Another example of how the lesson situation or context affects the observed scores, albeit to a limited extent, can be seen in Emma's observed scores for lessons C and D. These were the only two lessons in the dataset that were taught to lower school students (14-year-olds). Although the language level for these students in a bilingual stream is generally comparable to upper school regular students, i.e. CEFR B2 (Verspoor, De Bot, & Xu, 2015) the content and activities in the lessons are at a different cognitive level, which resulted in the use of different teaching strategies. We could say that these findings confirm that our instrument is indeed sensitive to the scaffolding strategies one of the teachers uses in different groups. Ultimately, a residual of 27.13% percent variance in our results cannot be explained by this model.

4.2. Number of observers needed: a decision study (D-study)

Classroom observations are labor intensive. In teacher training practice often only one observer evaluates. We therefore wanted to know how many observers are needed to achieve acceptable reliability for observations with this instrument.

The results of a G-study can be used to perform a decision study (D-study) (Shavelson &Webb,1991). The goal of a decision study is to gain insight into the extent to which the variance caused by observers and by interactions between observers and teachers affects the reliability of the observation scores. D-studies calculate the expected reliability and show the influence of systematic inconsistencies in the observed scores. Fig. 4 illustrates how many observers were needed to reach a reliable conclusion about the observed teaching strategies.

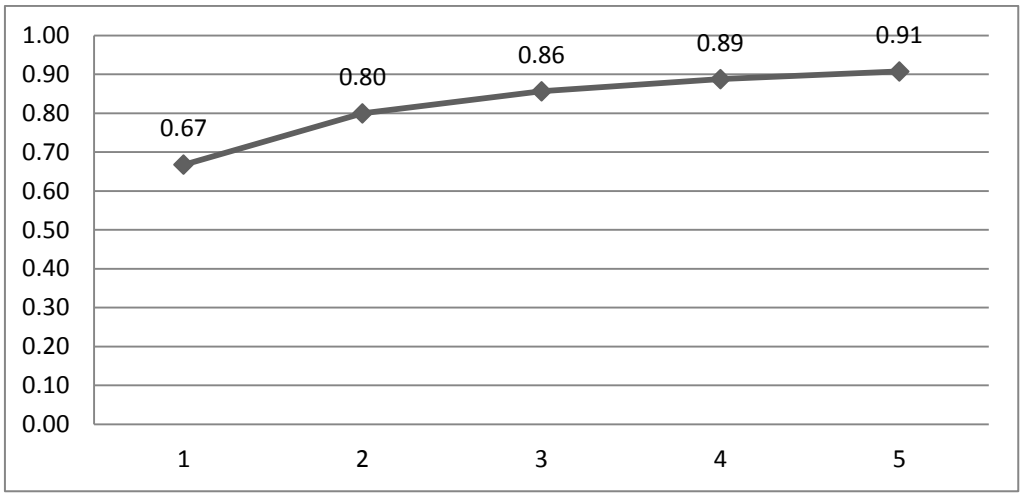


Figure 4 Results of the D-study relating to the reliability (ρ) for the number of observers

The reliability coefficient in D-study is expressed in terms of Spearman's r. The y-bar shows the increase in expected reliability with a larger number of observers. The low proportion of observer variance which we found in our G-study is also reflected in the results of our D-study: r ¼ 0.67 for one observer, r ¼ 0.80 for two observers. This indicates that only a very small proportion of unsystematic variance exists. These results also suggest that if we train observers to follow the procedure described in section 3.4, we would only need two observers for reliable observational results.

5. Conclusion

The aim of this study was to develop a reliable observation tool for teacher development purposes: a tool that can identify teachers' use of scaffolding strategies. We explored which scaffolding strategies a teacher is using in secondary EFL reading lessons, in which the focus is on fostering critical literacy in a foreign language program (Coyle et al., 2010; Llinares, Morton, & Whittaker, 2012). We wanted to observe reliably what a teacher already does in terms of asking questions, modeling, explaining, giving instructions, hints and feedback in a whole class setting. In the ERCOP observation tool we have operationalized scaffolding means for EFL reading comprehension. Through a stacked procedure we found 42 observation points that we can observe reliably. We established content validity through piloting the instrument and compiling a list of items that cause little debate. The results of our G-study show that a high percentage of the variance (51.7%) is caused by observed differences between teachers. From the low percentage of observer variance (21.2% in total) we can conclude that our instrument is not sensitive to personal observer biases. We are aware of the fact that whenever teachers are using the strategies listed in the protocol, there might be considerable differences in how they use them in the dynamic context of the EFL classroom.

The dataset for this G-study was not large enough to conduct separate G-studies and calculate reliability for the dimensions (instruction, feedback, explaining, modeling, hints, questioning). In this paper we used a sum score for scaffolding based on 41 observation points. In a larger G-study with more teachers and lessons it would also be possible to disentangle interaction effects between teachers, lessons, groups and observers.

Observing teacher behavior removes the self-report bias a researcher might find in interviews and surveys. If we want to compare classrooms or measure teacher development, a structured observation protocol is needed for stable observations.

We are aware that any observation scheme is selective in nature and that an event sampling scheme focuses on predetermined behavior, which is problematic regarding the dynamic nature of classroom. The concept of scaffolding is dynamic in essence and a structured observation protocol in the form of a unidimensional coding scheme is insufficient to capture the dynamic nature of scaffolding. The ERCOP instrument focuses on the range of teaching strategies, not on the sequence in which strategies are used. Although we view scaffolding as a complex interactive process, we contend that we first need to operationalize the individual steps of pedagogical sequences, and that we need to be able to observe these individual steps reliably. In this study observations of the behavior of pupils have been left out of the equation. In order to gain a better understanding of the teacher's role in the phase of contingency management and the process of fading, research into pupil-teacher interaction patterns should be considered in future research. Triangulation, for instance by supplementing observational data with self-reports, think-alouds or interviews, is recommended to better understand of the complexity and richness of language classrooms.

An essential first step for teachers in scaffolding the learning process is assessing the current level and grasping pertinent learning needs in EFL readers. We can draw an analogy for teacher education and professional development programs. If teachers need to learn to scaffold their learners' process of becoming skilled and critical EFL readers, a reliable diagnosis of the current use of their teaching strategies is an essential starting point.

The present study presents a practical and reliable instrument in which teaching strategies for scaffolding EFL reading comprehension have been made explicit. Scaffolding is a complex process. An observation list with concrete strategies that can be used for asking questions, modeling learning, and giving feedback, hints, instructions and explanations might help teachers to make their lessons more student-centered. Although the ERCOP tool was developed for secondary schools in the Dutch context, it might be used in different EFL teaching settings in order to start exploring and expanding teachers' strategic repertoires.

Acknowledgments

The authors wish to thank the reviewers for their valuable comments on a previous version of this manuscript. We also thank the observers, Marjolijn Verpoor, Marjon Tammenga-Helmantel, Jasmijn Bloemert and Rikkert van der Lans. Last but not least we want to express our gratitude to the participating teachers for their willingness, confidence and cooperation.

APPENDIX

ERCOP: EFL READING COMPREHENSION OBSERVATION PROTOCOL

Instruction

1 has good introductory hands-on tasks to build initial interest
2 relates text to students' background knowledge
3 uses a flexible lesson design to facilitate the student's reading process
4 makes the students active participants in the reading lessons
5 uses coherent lesson procedures
6 creates a supportive environment
7 monitors students while working independently on the reading task

Explaining

1 uses signposting of lesson goals during activities
2 teaches students to use strategies for deriving meaning of an unfamiliar word
3 teaches students to self-evaluate their reading
4 checks that students are aware of reading relevance
5 introduces key words critical to understanding important concepts
6 teaches useful phrases that are relevant for the task
7 explains the usefulness of the vocabulary for the students' needs

Hints

1 breaks down a complex reading task into sub-activities when necessary
2 helps students to make inferences from the context
3 encourages students to seek help from a peer
4 encourages more capable students to "learn by teaching"
5 encourages students to interact with equal peers
6 helps students to analyze the task
7 helps students to understand the characteristics (genre, register) of a text
8 helps students to clarify (passages from) the text
9 lets students create vivid mental images related to the text

Modeling

1 helps students to manage reading time efficiently
2 uses visual materials (e.g. video, pictures, graphic organizers)
3 helps students to give L2 output
4 uses the L2 for instruction for 75-100% the lesson

Feedback

1 uses repetition
2 uses clarification requests in L2
3 uses recasts in L2
4 uses explicit correction
5 gradually withdraws teacher support from the reading task
6 gives feedback on the students' learning process
7 gives positive praise when praise is due
8 allows the students to use the L1 in the lesson

Questioning

1 encourages students to generate questions about the text
2 builds on student-generated questions in the lesson
3 promotes student interpretations of the text
4 asks students for opinions about text content
5 uses thick questions
6 checks students' understanding of (passages from) the text

References

Assessment Reform Group (2002). *Assessment for Learning: 10 Principles. Research-based principles to guide classroom practice*. Retrieved from: http://www.aaia.org.uk/content/uploads/2010/06/Assessment-for-Learning-10-principles.pdf?v=796834e7a283

Cohen, L., Manion, L., Morrison, K., & Bell, R. (2011). *Research methods in education* (7th ed.). London etc: Routledge.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, U.K: Press syndicate of theUniversity of Cambridge.

Coyle, D., Hood, P., & Marsh, D. (2010). *Clil: Content and language integrated learning*. Cambridge: Cambridge University Press.

Dalton-Puffer, C. (2007). *Discourse in content and language integrated learning (CLIL) classrooms*. Amsterdam etc.: Benjamins.

De Graaff, R., Koopman, G. J., Anikina, Y., & Westhoff, G. (2007). An observation tool for effective L2 pedagogy in content and language integrated learning (CLIL). *International Journal of Bilingual Education & Bilingualism*, 10(5), 603-624. http://dx.doi.org/10.2167/beb462.0.

Derrick, D. J. (2015). Instrument reporting practices in second language research. *TESOL Quarterly*. http://dx.doi.org/10.1002/tesq.217. n/aen/a.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38(3), 181-199. http://dx.doi.org/10.3102/0013189X08331140.

Echevarria, J., Vogt, M. A., & Short, D. J. (2010). *Making content comprehensible for English learners: The SIOP model* (3rd ed.). Boston: Pearson Education.

Foley, J. (1994). Key concepts in ELT: Scaffolding. *ELT Journal*, 48(1), 101-102. http://dx.doi.org/10.1093/elt/48.1.101.

Fröhlich, M., Spada, N., & Allen, P. (1985). Differences in the communicative orientation of L2 classrooms. TESOL Quarterly, 19(1), 27-57. http://dx.doi.org/10.2307/3586771.

Gibbons, P. (2002). *Scaffolding language, scaffolding learning*. Portsmouth, NH: Heinemann.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York, N.Y., etc: Cambridge University Press.

Grabe, W., & Stoller, F. L. (2011). *Teaching and researching reading* (2nd ed.). Harlow: Pearson Education Limited.

Guilloteaux, M. J., & Dörnyei, Z. (2008). *Motivating language learners: A classroom-oriented investigation of the effects of motivational strategies on student motivation*. TESOL Quarterly, 42(1), 55-77. http://dx.doi.org/10.1002/j.1545-7249.2008.tb00207.x.

System 65 (2017)

Guthrie, J. T., Wigfield, A., & Perencevich, K. C. (2004). *Motivating reading comprehension: Concept-oriented reading instruction*. Mahwah, NJ etc: Laurence Erlbaum Associates.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London etc: Routledge.

Hattie, J. A. C. (2012). *Visible learning for teachers* (1st ed.). New York: Routledge.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough. *Educational Researcher*, 41(2), 56-64. http://dx.doi.org/10.3102/0013189X12437203.

Juan-Garau, M., & Jacob, K. (2015). Developing English learners' transcultural skills through content- and task-based lessons. *System*, 54, 55-68. http://dx.doi.org/10.1016/j.system.2015.04.017.

Lightbrown, P. M. (2014). *Focus on content-based language teaching*. Oxford: OUP.

Llinares, A., Morton, T., & Whittaker, R. (2012). *The roles of language in CLIL*. Cambridge, etc: Cambridge University Press.

Long, M. H., & Doughty, C. J. (2009). *The handbook of language teaching*. Chichester etc: Wiley-Blackwell.

Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach.* Amsterdam etc: Benjamins.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition*, 19(01), 37-66. http://dx.doi.org/10.1017/S0272263197001034.

Mackey, A., & Gass, S. M. (2012). *Research methods in second language acquisition : A practical guide*. Chichester, West Sussex; Malden, Mass.: Wiley-Blackwell.

Meijer, D., & Fasoglio, D. (2007). *Handreiking schoolexamen moderne vreemde talen havo/vwo*. Enschede: Stichting Leerplanontwikkeling.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50, 417. http://dx.doi.org/10.1111/0023-8333.00136.

Norris, J. M., & Ortega, L. (2006). *Synthesizing research on language learning and teaching*. Amsterdam: Benjamins.

Nuttall, C. (2005). *Teaching reading skills in a foreign language* (2nd ed.). Oxford: Macmillan.

OECD. (2016). *Education at a glance 2016: OECD indicators*. Paris: OECD Publishing. http://dx.doi.org/10.1787/eag-2016-en.

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching* (3rd ed.). Cambridge: CUP.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA etc: Sage Publications.

System 65 (2017)

Snow, M. A., Met, M., & Genesee, F. (1989). A conceptual framework for the integration of language and content in second/foreign language instruction. *TESOL Quarterly*, 23(2), 201-217. http://dx.doi.org/10.2307/3587333.

Spada, N., & Lyster, R. (1997). Macroscopic and microscopic views of L2 classrooms. *TESOL Quarterly*, 31(4), 787-795. http://dx.doi.org/10.2307/3587763.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature. *Language Learning*, 60(2), 263-308. http://dx.doi.org/10.1111/j.1467-9922.2010.00562.x.

Stichting Leerplan Ontwikkeling. (2013). *Leerplan in beeld*. Retrieved from http://leerplaninbeeld.slo.nl/havo_vwo_bovenbouw/moderne-vreemde-talen/engels/engels-po-havo-vwo/.

Tharp, R. G., & Gallimore, R. (1988). *Rousing minds to life: Teaching, learning and schooling in social context*. Cambridge: Cambridge University Press.

Van Geert, P., & Steenbeek, H. (2005). The dynamics of scaffolding. *New Ideas in Psychology*, 23(3), 115e128. http://dx.doi.org/10.1016/j.newideapsych.2006.05.003.

Van Gorp, K., & Van den Branden, K. (2015). Teachers, pupils and tasks: The genesis of dynamic learning opportunities. *System*, 54, 28-39. http://dx.doi.org/10.1016/j.system.2015.04.018.

Van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and an application of an assessment instrument. *Educational Research*, 49(2), 127-152.

Van de Grift, W. (2009). Reliability and validity in measuring the value added of schools. *School Effectiveness and School Improvement*, 20(2), 269-285. http://dx.doi.org/10.1080/09243450902883946.

Van de Grift, W., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation*, 43, 150-159. http://dx.doi.org/10.1016/j.stueduc.2014.09.003.

Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. Educational Psychology Research, 22(3),271-296. http://dx.doi.org/10.1007/s10648-010-9127-6.

Verspoor, M., De Bot, K., & Xu, X. (2015). The effects of English bilingual education in the Netherlands. *Journal of Immersion and Content-Based Language Education*, 3(1), 4. http://dx.doi.org/10.1075/jicb.3.1.01ver.

Walqui, A. (2006). Scaffolding instruction for English language learners: A conceptual framework. *International Journal of Bilingual Education and Bilingualism*, 9(2), 159-180. http://dx.doi.org/10.1080/13670050608668639.

Walqui, A., & Van Lier, L. (2010). *Scaffolding the academic success of adolescent English language learners*. San Fransisco: WestEd.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89-100. http://dx.doi.org/